ГЕОМАГНЕТИЗМ И АЭРОНОМИЯ, 2023, том 63, № 2, с. 190–201

УДК 537.591.5

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ПРИ ИССЛЕДОВАНИИ ВАЖНОСТИ ВХОДНЫХ ПРИЗНАКОВ В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ ГЕОМАГНИТНОГО ИНДЕКСА DST

© 2023 г. Р. Д. Владимиров[1], *, В. Р. Широкий[1], **, И. Н. Мягкова[1], *, О. Г. Баринов[1], ****, С. А. Доленко[1], *****

[1]*Научно-исследовательский институт ядерной физики им. Д.В. Скобельцына Московского государственного университета им. М.В. Ломоносова (НИИЯФ МГУ), Москва, Россия*

**e-mail: vladimirov.rd16@physics.msu.ru*

***e-mail: shiroky@srd.sinp.msu.ru*

****e-mail: irina@srd.sinp.msu.ru*

*****e-mail: obar@sinp.msu.ru*

******e-mail: dolenko@srd.sinp.msu.ru*

Поступила в редакцию 02.03.2022 г.
После доработки 11.11.2022 г.
Принята к публикации 28.11.2022 г.

Одним из перспективных подходов к прогнозированию значений геомагнитных индексов является использование методов машинного обучения. Однако для эффективного использования таких методов необходим отбор существенных входных признаков задачи с целью уменьшения ее входной размерности. В данной работе рассматривается алгоритм получения наиболее эффективной модели прогнозирования, основанный на понижении входной размерности данных путем постепенного отбрасывания входных признаков на основе следующих методов машинного обучения: линейная регрессия, градиентный бустинг, искусственная нейронная сеть типа многослойный персептрон. Проводится сравнение эффективности перечисленных методов; рассматриваются направления дальнейшего развития работ.

DOI: 10.31857/S0016794022100224, **EDN:** DLYMRJ

1. ВВЕДЕНИЕ

Одним из наиболее общеупотребительных геомагнитных индексов, содержащих информацию о планетарных возмущениях во время геомагнитных бурь, является *Dst*-индекс, который представляет собой меру изменения поля из-за кольцевых токов, возникающих в магнитосфере во время магнитных бурь (*Disturbance Storm Time*).

Данный индекс был введен М. Сугиурой в 1964 г. [Sugiura, 1964]. (Данные доступны с начала 1957 г.) Согласно его работе, *Dst*-индекс вычисляется как средняя в часовом интервале величина возмущения горизонтальной составляющей напряженности магнитного поля Земли, отсчитываемого от спокойного уровня, определенная по данным четырех низкоширотных обсерваторий, равномерно распределенных по географической долготе: Какиока, Гонолулу, Сан-Хуан, Херманус.

Если *Dst*-индекс вычисляется непрерывно как функция времени (UT), то его вариация отображает возникновение магнитных бурь и их интенсивность. Таким образом, *Dst*-индекс представляет собой количественное измерение геомагнитного возмущения, и его можно сопоставлять с солнечными и геофизическими параметрами, а прогнозирование амплитуды *Dst*-индекса позволяет оценивать не только время начала, но и мощность геомагнитного возмущения.

Как известно, основным физическим явлением, влияющим на магнитосферу Земли, являются потоки ионизованных частиц от Солнца – солнечный ветер (СВ) [Akasofu et al., 1972]. Источниками возмущений магнитосферы Земли являются корональные выбросы массы (КВМ), достигающие орбиты Земли, и высокоскоростные потоки СВ [Ермолаев Ю.И., Ермолаев М.Ю., 2009]. Необходимым и достаточным условием возникновения магнитных бурь на Земле является относительно большая по амплитуде и продолжительности существования южная ($Bz < 0$) компонента межпланетного магнитного поля (ММП). Это делает магнитосферу “открытой” для поступления в нее энергии солнечного ветра.

Геомагнитные бури (или геомагнитные возмущения, поскольку бурями принято называть возмущения, имеющие амплитуду выше определенной) представляют собой один из наиболее существенных факторов космической погоды, который с развитием мировой космической отрасли становится все более важным [Pulkkinen, 2007; Schrijver et al., 2015].

В прикладном аспекте прогнозирование геомагнитных возмущений представляет интерес, так как магнитные бури могут стать причиной нарушений в работе телеграфных линий и радиосвязи, трубопроводов, линий электропередач и энергосетей [Лазутин, 2012]. Помимо того, магнитные бури опосредованно оказывают существенное влияние на состояние околоземного космического пространства (ОКП) за счет того, что после примерно половины магнитных бурь на порядок и более возрастает поток релятивистских электронов внешнего радиационного пояса Земли (РПЗ) [Kataoka R., Miyoshi Y., 2008; Мягкова и др., 2013]. Экстремальные потоки электронов внешнего РПЗ могут привести к сбоям в электронных микросхемах аппаратуры, находящейся на борту космических аппаратов (например, [Белов и др., 2004]). Поэтому прогнозирование геомагнитных возмущений представляется весьма актуальной задачей.

С точки зрения фундаментальной науки исследование механизмов переноса энергии из солнечного ветра в магнитосферу Земли, и, как следствие этого переноса, возникновения магнитосферных возмущений также представляет интерес, поскольку это один из наиболее актуальных вопросов физики солнечно-земных связей.

Dst-индекс, как и большинство геомагнитных индексов, имеет долговременную историю наблюдения, что делает возможными как статистические исследования связи геомагнитной активности с процессами в межпланетном пространстве, солнечном ветре и магнитосфере Земли [O'Brien, McPherron, 2000; Pallochia et al., 2006; Podladchikova, Petrukovich, 2012; Patra et al., 2011], большинство из которых основано на формуле Бёртона [Burton et al., 1975], так и машинное обучение [Lindsay et al., 1999; Barkhatov et al., 2000; Bortnik et al., 2018; Wu, Lundstedt, 1997; Lazzús et al., 2017; Revallo et al., 2014]. Сравнение качества прогнозирования разными моделями выполнено в работе [Amata et al., 2008].

Авторами настоящей работы – сотрудниками лаборатории адаптивных методов обработки данных НИИЯФ МГУ еще в 2005 г. было показано, что лучшее качество прогноза *Dst*-индекса достигается при построении нейросетевой модели, использующей в качестве входных данных как историю *Dst* индекса, так и параметры СВ (скорость) и ММП (компонента *Bz*) [Dolenko et al., 2005]. В более поздних работах [Широкий, 2015; Myagkova et al., 2017] каждый пример содержал среднечасовые значения нескольких основных параметров СВ и ММП и часовые значения самого прогнозируемого параметра – индекса *Dst* – с топологическим вложением (учетом предыдущих значений) временно́го ряда на 24 ч, что позволило улучшить качество прогноза. Использование подобного подхода стало возможным только в последние годы, когда накопились достаточно длинные однородные временны́е ряды спутниковых измерений параметров СВ и ММП, полученные в ходе эксперимента на КА ACE, передающем данные с октября 1997 г. В работе [Ефиторов и др., 2018] было получено, что при прогнозировании амплитуды *Dst*-индекса при помощи ИНС двух разных типов – классических персептронов и рекуррентных сетей типа LSTM, а также комитетов прогнозирующих моделей, наилучшие результаты достигаются при использовании гетерогенных комитетов на основе ИНС обоих типов.

2. ПОСТАНОВКА ЗАДАЧИ

Целью данной работы является совершенствование алгоритмов прогнозирования значения геомагнитного индекса *Dst* с горизонтом от 1 до 6 ч.

В работе [Мягкова и др., 2021] авторами настоящей работы было произведено сравнение показателей качества прогнозирования индекса *Dst* с горизонтом от 1 до 6 ч с помощью трех методов машинного обучения–случайного леса [Breiman, 2001], градиентного бустинга [Friedman, 2002] и искусственных нейронных сетей типа многослойный персептрон [Haykin, 1998]. Было показано, что наилучшие результаты среди трех перечисленных методов обеспечивает градиентный бустинг.

Однако в указанной работе размерность входных данных была достаточно велика и составляла около 130. Это было связано с тем, что для всех входных физических величин (описанных в следующем разделе) для каждого примера в качестве входных признаков использовались все их предыдущие значения с задержкой от 1 до 24 ч. Между тем, в пространствах высокой размерности имеет место эффект концентрации меры, одним из следствий которого является тот факт, что малая окрестность медианного уровня любой функции, непрерывной на многомерной сфере, содержит почти всю сферу. Поэтому с точки зрения наблюдателя, измеряющего значения такой функции, она представляется практически постоянной [Зорич, 2014]. Иными словами, любой нелинейный многомерный предиктор должен давать прогноз, близкий к тривиальному (инерционному), смысл которого в контексте прогнозирования времен-

ны́х рядов – равенство прогнозируемого значения последнему известному значению прогнозируемого временно́го ряда. Ясно, что практическая ценность такого тривиального прогноза равна нулю, поэтому любой прогноз в пространствах высокой размерности следует сравнивать по его статистическим показателям с тривиальным. При этом есть основания ожидать, что с уменьшением размерности пространства входных признаков и связанным с этим ослаблением эффекта концентрации меры качество прогноза, выполненного методом машинного обучения, и его преимущество над тривиальным может дополнительно увеличиться.

Поэтому в контексте решения задачи прогнозирования понижение размерности входных данных представляется весьма существенным. При этом такое понижение размерности может быть достигнуто либо с помощью алгоритмов преобразования признаков, либо с помощью алгоритмов отбора наиболее существенных признаков. В последнем случае, помимо собственно понижения размерности, представляет интерес анализ множества признаков, отобранных в качестве наиболее существенных, что может дать информацию как об относительной существенности значений тех или иных физических величин, так и о необходимой глубине учета предыстории.

В настоящей работе была поставлена задача – провести сравнительный анализ результатов применения алгоритма получения наиболее эффективной модели прогнозирования путем постепенного отбрасывания входных признаков на основе следующих методов машинного обучения: линейная регрессия, градиентный бустинг, искусственная нейронная сеть типа многослойный персептрон. Алгоритм описан ниже в разделе 4.

3. ИСХОДНЫЕ ДАННЫЕ

Поскольку процессы, происходящие в земной магнитосфере, солнечном ветре и гелиосфере, взаимосвязаны, прогнозирование основных факторов космической погоды в системе “Солнце–гелиосфера–солнечный ветер–магнитосфера” основываются на данных экспериментов на космических аппаратах и наземных геофизических станциях, получаемых в режиме реального времени. При краткосрочном прогнозировании нам необходима оперативная информация о значениях параметров СВ и ММП и самого геомагнитного индекса. Входными данными, используемыми нами для прогнозирования *Dst*-индекса, являются параметры плазмы СВ и параметры ММП, измеренные в точке Лагранжа *L1* между Солнцем и Землей, полученные в эксперименте на борту КА ACE (Advanced Composition Explorer) (https://izw1.caltech.edu/cgi-bin/dib/rundibviewbr/ACE/ASC/DATA/browse-data?ACE_BROWSE.HDF!hdfref;tag=1962,ref=3,s=0).

Помимо данных о параметрах ММП и СВ, использовались значения индекса *Dst*, полученные с сайта Всемирного центра данных по геомагнетизму в Киото (https://wdc.kugi.kyoto-u.ac.jp/dstae/index.html).

При построении прогнозирующих моделей нами использовались следующие данные – временны́е ряды (ВР) среднечасовых (часовых) значений следующих физических величин:

а) Параметры СВ в точке Лагранжа *L1* между Землей и Солнцем:

- Скорость СВ *V* (км/с)
- Плотность протонов в СВ *Np* ($см^{-3}$)

б) Параметры вектора ММП в той же точке Лагранжа *L1* в системе *GSM* (нТл):

- *By*, *Bz* (*y*- и *z*-компоненты ММП)
- B_magn (модуль ММП)

в) Геомагнитный индекс *Dst* (нТл).

Помимо этого, для учета суточных и годовых вариаций *Dst*-индекса во временно́й ряд включались значения, характеризующие привязку этого примера к определенным фазам суточного и годового циклов. Для этого использовались значения косинуса и синуса времени с суточным и годовым периодами. Гармоническая функция времени позволяет учесть непрерывность и периодичность используемых временны́х зависимостей, а одновременное применение двух гармонических зависимостей с одним периодом, сдвинутых по фазе (синус, косинус), делает привязку к определенной фазе цикла однозначной [Ефиторов и др., 2018].

Для учета предыстории использовалось погружение (топологическое вложение) всех ВР на глубину в 23 ч, т.е. на вход алгоритма, помимо текущих значений всех входных величин, подавались их предыдущие значения за 1, 2, 3, ... 23 ч до текущего. Такая глубина погружения была выбрана, поскольку видится достаточной для работы с данными, имеющими часовое временно́е разрешение.

Таким образом, полная входная размерность данных составляла $6 \times 24 + 4 = 148$.

Отметим также, что нами использовались не предобработанные и очищенные данные 2-го уровня (*Level 2 Data*), предназначенные для научных исследований, а оперативные данные (*Browse Data*), поскольку разрабатываемая система прогнозирования *Dst*-индекса предназначена для использования в режиме онлайн, в котором качество получаемых данных соответствует оперативным данным. Поэтому машинное обучение следует проводить для работы с данными такого качества.

Рис. 1. Общая схема алгоритма оценки важности признаков.

Также данные *Level 2* имеют значительно большее число пропусков, что затрудняет применение методов машинного обучения, в особенности с учетом погружения ВР. Пропуски в данных размером 12 ч и менее заполнялись линейной интерполяцией по двум крайним точкам; оставшиеся примеры с пропусками удалялись из набора.

В настоящей работе использовался массив данных с ноября 1997 г. (со времени начала поступления данных с КА *ACE*) по декабрь 2021 г. включительно (всего около 212000 примеров). Имеющийся массив был разбит на обучающую выборку и тестовый набор данных. Обучающая выборка использовалась для обучения алгоритмов (подстройки настраиваемых параметров моделей), тестовый набор – для независимой оценки результатов обучения. Для ИНС обучающая выборка дополнительно разбивалась на тренировочный и валидационный наборы данных. Тренировочный набор использовался для подстройки весов при обучении ИНС, валидационный набор – для периодической проверки в процессе обучения с целью предотвращения переучивания. Для тренировочного и валидационного набора использовались данные с ноября 1997 г. по декабрь 2016 г. включительно, тренировочный и валидационный наборы делились случайным образом в соотношении 80% к 20%. Тестовый набор составляли данные с 2017 по 2021 г.

4. ОПИСАНИЕ АЛГОРИТМА ОТБОРА СУЩЕСТВЕННЫХ ВХОДНЫХ ПРИЗНАКОВ

Схема алгоритма оценки важности признаков приведена на рис. 1.

Для работы алгоритма оценки важности необходим базовый метод построения модели прогнозирования, получающий на вход любое количество признаков от 1 до N (плюс 4 гармонические функции), и обучающий набор данных.

Алгоритм обучается следующим образом.

Шаг 1. На полном наборе из N признаков обучаем базовую модель и оцениваем качество ее работы. В дальнейшем эта модель используется в качестве референсной, наряду с тривиальной моделью (прогнозируемое значение = последнему известному).

Шаг 2. Поочередно отбрасываем признаки по одному. Тренируем N моделей с $N - 1$ входным признаком каждая, оцениваем качество каждой из них.

Шаг 3. Наименее важным считаем признак, при удалении которого получается наилучшая модель. Окончательно удаляем его из множества входных признаков.

Шаг 4. Если в наборе осталось более 1 признака, присваиваем $N = N - 1$ и возвращаемся к Шагу 2.

По окончании работы алгоритма все признаки оказываются ранжированными в порядке отбрасывания, т.е. по возрастанию важности. Оптимальным считается набор входных признаков, которому соответствует наилучший показатель качества построенной на нем модели.

В данной работе рассматривались два варианта базового алгоритма построения модели: линейная регрессия (ЛР) и градиентный бустинг (ГБ).

ЛР имеет существенно меньшую вычислительную стоимость, однако линейность этого алгоритма может приводить к худшему качеству прогнозирования.

Алгоритм ГБ представляет собой ансамблирование моделей деревьев принятия решений, каждая из которых (в соответствии с подходом бустинга) исправляет ошибки предыдущей, а коэффициенты моделей подбираются алгоритмом градиентного спуска. Модель деревьев принятия решений [Breiman et al., 1983] представляет собой кусочно-линейный аппроксиматор и способна с заданной точностью описывать нелинейные аппроксимируемые функции, если это позволяют как размер и представительность обучающей выборки, так и параметры алгоритма, в особенности глубина построенного графа принятия решений, увеличение которой обеспечивает более плавную аппроксимацию. При ансамблировании строится множество таких графов, и каждый последующий позволяет аппроксимировать все более сложную и нелинейную функцию. По этой причине, несмотря на линейность отдельного узла дерева, алгоритм градиентного бустинга в целом представляет собой алгоритм нелинейной (кусочно-линейной) аппроксимации.

На данном этапе работ многослойный персептрон (МСП) в качестве базового алгоритма не рассматривался ввиду высокой вычислительной стоимости построения модели (полная реализация алгоритма отбора требует на имеющихся вычислительных мощностях около двух месяцев непрерывного счета, что кратно превышает вычислительную стоимость для ГБ).

Таким образом, было получено два лучших набора входных признаков: отобранный с помощью ЛР (на нем была получена лучшая ЛР-модель) и отобранный с помощью ГБ (на нем была получена лучшая ГБ-модель).

Третий лучший набор входных признаков получался следующим образом. Рассматривались последовательности наборов входных признаков, полученные с помощью ЛР в диапазоне от 3 до 31 входного признака и отдельно с помощью ГБ в том же диапазоне. На каждом из этих наборов обучалось по 5 МСП с различными инициализациями весов; ответы этих 5 сетей усреднялись. Наилучшим для МСП считался набор входных признаков, для которого такой одноранговый комитет из 5 МСП дал наилучший показатель качества. Этот же набор признаков считался наиболее физически содержательным при анализе отобранных признаков (см. ниже), поскольку МСП в силу своей нелинейности и свойства универсальной аппроксимации отбирает меньшее количество признаков, чем ЛР или ГБ.

Таким образом, для каждого горизонта прогнозирования от 1 до 6 ч сравнивались между собой результаты следующих моделей:

- Тривиальная модель;
- ЛР на лучшем для ЛР наборе признаков, отобранном с помощью алгоритма (рис. 1);
- ГБ на лучшем для ГБ наборе признаков, отобранном с помощью алгоритма (рис. 1);
- МСП на лучшем для МСП наборе признаков, отобранном с помощью описанной выше процедуры из наборов, полученных в процессе работы алгоритма (рис. 1) для ЛР и ГБ.

5. РЕЗУЛЬТАТЫ

Изложим здесь основные полученные результаты. Как критерии качества полученных моделей мы везде используем среднеквадратичное отклонение (СКО) и коэффициент детерминации R^2, рассчитанные на тестовом наборе данных (2017–2021 гг.).

Зависимости СКО от количества входных признаков модели носят схожий характер для всех горизонтов прогнозирования и для обоих базовых алгоритмов (ЛР и ГБ). На рис. 2 приведена такая зависимость для прогноза на 1 ч с помощью ЛР. На рис. 3 – с помощью ГБ. Видно, что в целом зависимость является выраженной весьма слабо, а СКО показывает выраженный монотонный рост лишь при общем количестве признаков менее 15–20. Однозначный вывод, который можно сделать из наблюдаемых зависимостей – количество входных признаков для обоих базовых алгоритмов (ЛР и ГБ) может быть кратно уменьшено без потери или даже с некоторым повышением качества прогнозирования.

С точки зрения анализа физических взаимосвязей наибольший интерес представляют минимальные (среди трех алгоритмов) по количеству признаков конфигурации, отобранные описанным выше способом с помощью МСП на основе

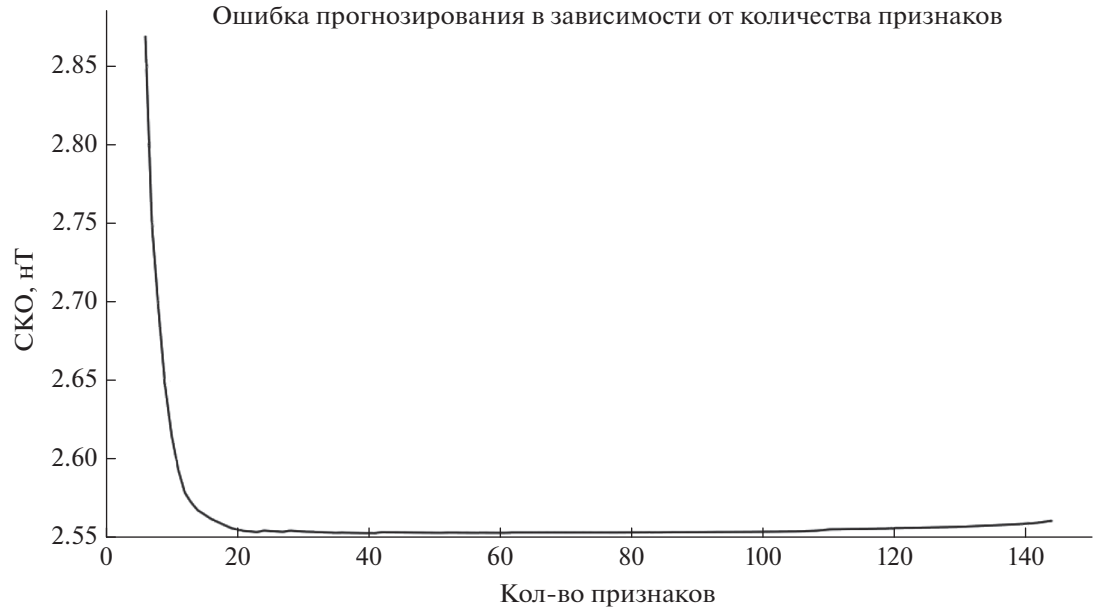

Рис. 2. Зависимость среднеквадратичного отклонения прогноза на 1 ч с помощью алгоритма линейной регрессии на тестовом наборе от числа входных признаков.

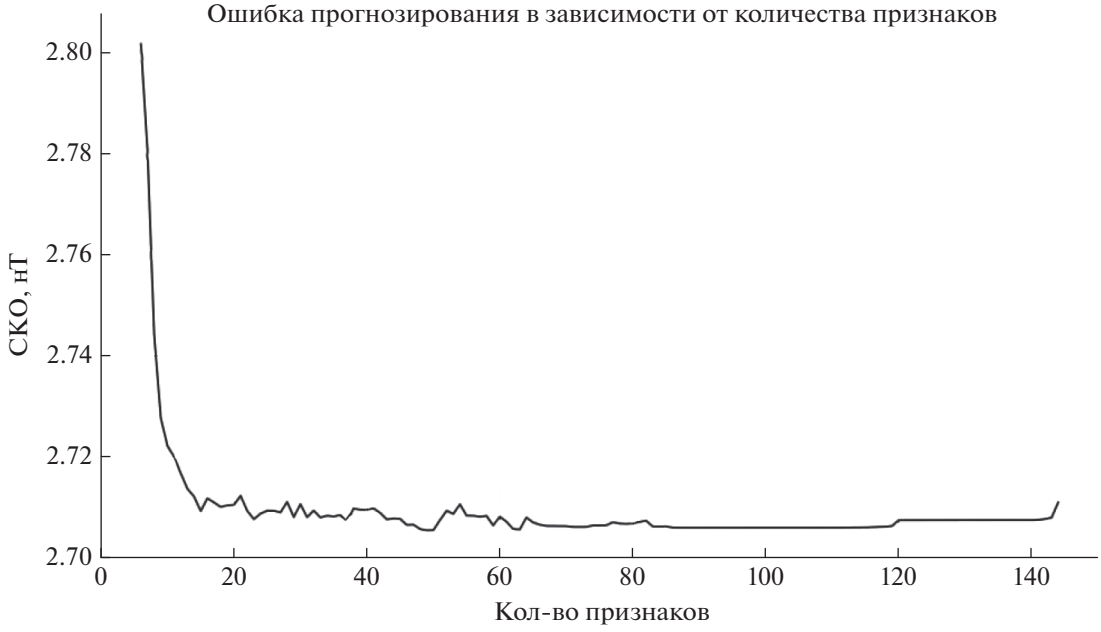

Рис. 3. Зависимость среднеквадратичного отклонения прогноза на 1 ч с помощью алгоритма градиентного бустинга на тестовом наборе от числа входных признаков.

применения алгоритма поэтапного отбрасывания признаков на базе ЛР и ГБ. На двух следующих рисунках отмечены входные признаки, отобранные в качестве существенных на базе ЛР (рис. 4) и на базе ГБ (рис. 5). В столбцах – входные физические величины и горизонт прогноза в часах (от 1 до 6); в строках – задержка используемого значения относительно момента прогнозирования (от 0 до 23).

Сравнение рис. 4 и 5 показывает, что отбираемые комплекты признаков несколько различаются; для логического завершения работы необходимо провести отбор существенных признаков с поочередным удалением на базе самого МСП.

Признак	*Dst*						*By* GSM						*Bz* GSM						*B*_magn						SW_spd						H_den_SWP					
Горизонт	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
0	X	X	X	X	X	X	o	o	X	X	X	X	X	X	X	X	X	X	o	X	X	X	X	X	X	o	o	X	X	X	X	X	X	X	X	o
1	X	X	X	X	X	X	o	X	o	o	o	o	X	X	X	X	o	X	X	o	o	X	o	o	X	X	X	o	o	o	X	X	X	X	X	X
2	X	X	X	X	X	o	o	o	o	o	o	o	X	X	X	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
3	o	o	o	o	o	X	o	o	o	o	o	o	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
4	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
5	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
6	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
7	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
8	o	o	X	o	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
9	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o
10	X	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	X	o	o	o	o	o	o
11	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
12	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	X	o	o	o	o	o	o	o	o
13	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
14	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
15	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
16	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
17	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o
18	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
19	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	X	o	o	o	o	o	o	X
20	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o
21	o	X	X	X	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o
22	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	X	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
23	o	X	X	X	X	o	o	o	o	o	o	o	o	X	X	X	o	o	o	o	o	X	X	X	o	o	o	X	X	X	o	o	X	X	X	X

Рис. 4. Существенные входные признаки, отобранные с помощью МСП на основе применения алгоритма поэтапного отбрасывания признаков на базе ЛР (отмечены цветом и символом “X”). В столбцах – входные физические величины и горизонт прогноза от 1 до 6 ч. В строках – задержка в часах относительно момента прогнозирования.

Однако можно заметить, что примененный адаптивный метод отбора, не имеющий под собой прямой физически обусловленной модели, определил как наименее существенную из рассмотренных физическую величину *By* GSM. В число наиболее существенных входят значения самого прогнозируемого индекса *Dst*, а также значения *Bz* GSM и плотности с малыми задержками относительно момента прогнозирования (от 0 до 3 ч).

На рис. 6–8 показаны примеры результатов прогнозирования различными алгоритмами (ЛР, ГБ, МСП на лучших наборах признаков, а также тривиальная модель) на 1–3 ч и истинный временно́й ход индекса *Dst* для умеренной магнитной бури 27–29 августа 2021 г. Видно, что все алгоритмы прогнозирования работают существенно лучше, чем тривиальная модель, а качество прогноза заметно деградирует с увеличением горизонта. Статистические показатели использованных моделей на тестовом наборе в целом приведены в табл. 1.

Из результатов в табл. 1 видно, что наилучшее качество прогнозирования из всех рассмотренных моделей обеспечивает градиентный бустинг. Это в целом соответствует результатам, полученным нами ранее без отбора существенных признаков в работе [Мягкова и др., 2021], в которой ГБ также показал более высокие результаты, чем МСП. Однако для полностью адекватного сравнения методов необходимо, чтобы отбор существенных признаков каждый раз осуществлялся на базе того же метода машинного обучения, с помощью которого осуществляется прогнозирование (в нынешней постановке для МСП это условие не выполнено). Это планируется сделать при дальнейшем развитии данного исследования.

Можно также обратить внимание на тот факт, что для ГБ и ЛР отбор существенных признаков в большинстве случаев не приводит к улучшению качества модели, позволяя, однако, получить модель с практически теми же статистическими показателями при кратно меньшем количестве

Признак	*Dst*						*By* GSM						*Bz* GSM						*B*_magn						SW_spd						H_den_SWP					
Горизонт	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
0	X	X	X	X	X	X	o	o	X	o	X	o	X	X	X	X	X	X	o	X	X	X	X	X	o	X	X	X	X	X	X	X	X	X	o	o
1	o	o	X	X	X	X	o	o	o	X	o	X	X	X	X	X	X	X	o	o	X	X	o	o	X	X	X	o	o	o	X	X	X	X	X	X
2	o	o	o	o	X	o	o	o	o	o	o	o	o	X	X	o	o	o	o	o	X	X	X	X	o	o	X	o	X	o	X	X	X	X	o	o
3	o	X	X	o	o	o	o	o	o	o	o	o	X	X	X	X	o	X	o	o	o	X	X	o	o	o	o	o	o	X	o	o	o	X	o	X
4	o	o	o	X	X	X	o	o	o	o	o	o	o	o	o	o	o	X	o	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
5	o	X	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o
6	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	X	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o
7	o	X	o	o	X	X	o	o	X	o	o	o	o	o	X	o	X	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o
8	o	o	X	X	X	X	o	X	o	X	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
9	o	o	o	o	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o
10	o	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o
11	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	X	o	X	o	o	o	o	o	o	o	o	o	o	o	o
12	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o
13	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
14	o	o	o	o	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
15	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
16	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
17	o	o	X	o	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
18	o	o	o	o	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o
19	o	o	X	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	X	o	o	o	o	o	o
20	o	o	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
21	o	X	X	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o
22	X	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	X	o	o	o	o	o	o	o	o	o	o	o	o	o	o
23	o	o	X	X	o	o	o	o	X	o	X	o	o	o	o	o	o	o	o	o	X	X	X	X	o	o	o	o	X	o	o	o	o	o	o	o

Рис. 5. Существенные входные признаки, отобранные с помощью МСП на основе применения алгоритма поэтапного отбрасывания признаков на базе ГБ (отмечены цветом и символом “X”). В столбцах – входные физические величины и горизонт прогноза от 1 до 6 ч. В строках – задержка в часах относительно момента прогнозирования.

входных признаков (аналогичный вывод уже был сделан выше на основе зависимости статистических показателей от количества признаков при отборе, см. рис. 2, рис. 3). Для МСП ситуация иная: даже при использовании в качестве базового алгоритма при отборе признаков метода, отличного от самого МСП (ГБ или ЛР) удается получить модели, качество которых выше, чем качество МСП-моделей, построенных на полном наборе признаков (не говоря о вычислительной стоимости такой модели).

Объяснить наблюдаемый эффект можно следующим образом. Разные методы машинного обучения по-разному учитывают входные данные, что непосредственно влияет на отбор, и что хорошо заметно по результатам, приведенным в нижней части таблицы 1 для коэффициента детерминации. В частности, МСП всегда использует все входные признаки, из-за чего при наличии в наборе избыточных признаков является в наибольшей степени подверженным как переучиванию, так и эффекту концентрации меры. Позитивный эффект от понижения входной размерности наблюдается для всех горизонтов прогноза: при использовании только входных признаков, отобранных с помощью ЛР, коэффициент детерминации моделей МСП растет по сравнению с МСП, натренированными на полном наборе признаков. Иная ситуация наблюдается для моделей на основе ЛР и ГБ. ЛР содержит многократно меньшее количество подстраиваемых параметров, чем МСП, и оказывается в состоянии “отключить” малосущественные признаки, устанавливая для них в процессе обучения малые величины регрессионных коэффициентов. Алгоритм ГБ отбирает наиболее существенные признаки явным образом в процессе построения каждого составляющего модель дерева решений. Результатом этого является тот факт, что внешний по отношению к алгоритму отбор по итогу мало влияет на результаты этих двух алгоритмов: разница между коэффициентом детерминации на полном и на “лучшем” наборе данных для обоих

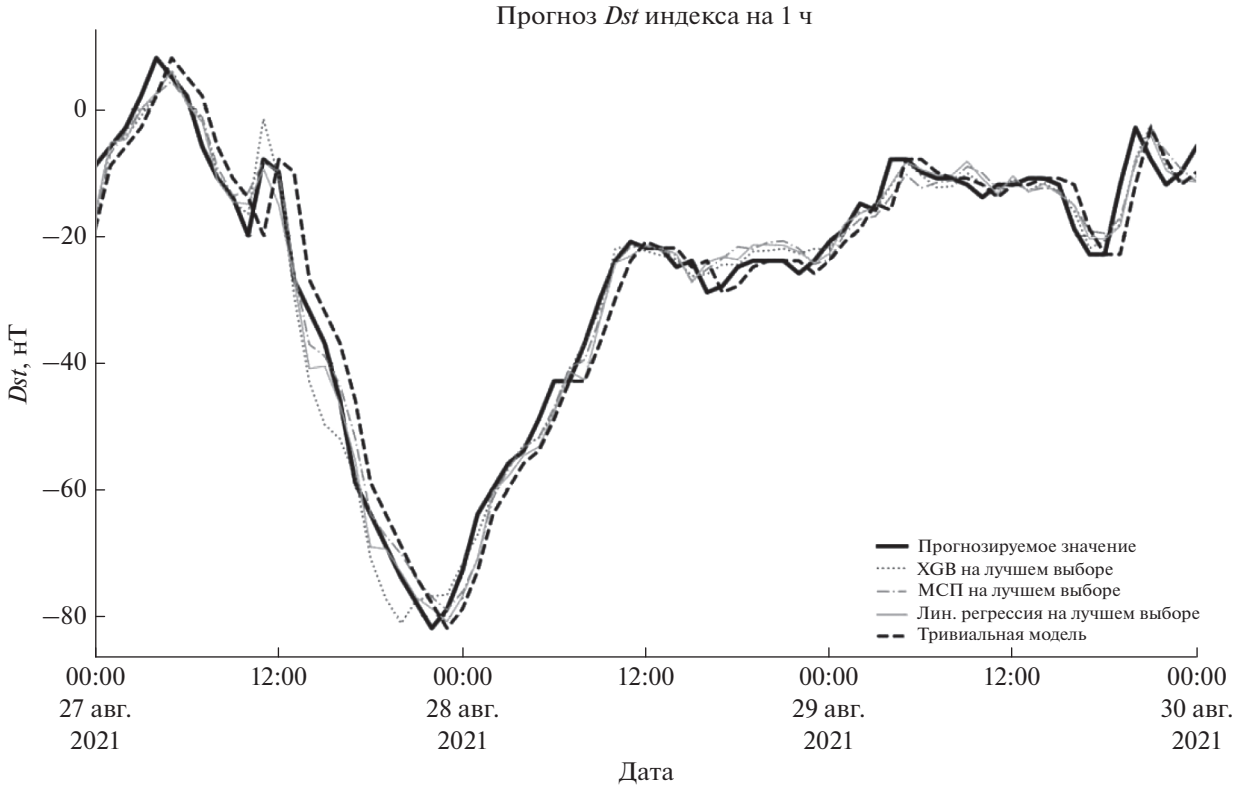

Рис. 6. Пример прогноза индекса *Dst* на 1 ч с помощью разных рассматриваемых моделей, и истинный временно́й ход индекса *Dst*.

Рис. 7. Пример прогноза индекса *Dst* на 2 ч с помощью разных рассматриваемых моделей, и истинный временно́й ход индекса *Dst*.

Рис. 8. Пример прогноза индекса *Dst* на 3 ч с помощью разных рассматриваемых моделей, и истинный временнóй ход индекса *Dst*.

Таблица 1. Сравнение статистических показателей разных моделей на тестовом наборе данных (2017–2021 гг.). Вверху – среднеквадратичное отклонение (СКО) в нТл; внизу – коэффициент детерминации (R^2)

СКО	Модель, набор признаков							
горизонт	ГБ полный	ГБ лучший	МСП полный	МСП лучш. ГБ	МСП лучш. ЛР	ЛР полный	ЛР лучший	тривиальная модель
1	**2.460**	2.469	2.912	2.781	2.717	2.615	2.621	3.384
2	**3.810**	3.845	4.254	4.158	4.135	4.136	4.152	5.468
3	**4.963**	4.970	5.315	5.326	5.307	5.323	5.329	6.856
4	**5.909**	5.915	6.196	6.215	6.208	6.261	6.266	7.828
5	6.579	**6.565**	6.818	6.872	6.868	6.928	6.935	8.568
6	**7.049**	7.054	7.243	7.333	7.320	7.382	7.403	9.181
R^2	Модель, набор признаков							
горизонт	ГБ полный	ГБ лучший	МСП полный	МСП лучш. ГБ	МСП лучш. ЛР	ЛР полный	ЛР лучший	тривиальная модель
1	**0.962**	0.961	0.946	0.951	0.956	0.957	0.956	0.927
2	**0.908**	0.906	0.885	0.890	0.897	0.891	0.890	0.810
3	**0.843**	**0.843**	0.820	0.819	0.830	0.820	0.819	0.701
4	**0.778**	0.777	0.756	0.754	0.768	0.750	0.750	0.610
5	0.724	**0.726**	0.704	0.699	0.716	0.694	0.694	0.533
6	**0.684**	0.683	0.666	0.658	0.677	0.653	0.651	0.463

Примечание. ГБ – градиентный бустинг, МСП – многослойный персептрон, ЛР – линейная регрессия; полный – полный набор входных признаков без отбора; лучший – лучший отобранный набор входных признаков при использовании указанного алгоритма в качестве базового.

алгоритмов не превышает 0.002, что не является значимым.

Наблюдаемый для МСП позитивный эффект понижения входной размерности дает основания полагать, что при использовании МСП в качестве базовой модели при отборе могут быть найдены множества существенных признаков, обеспечивающих наивысшее качество прогнозирования. Несмотря на высокую вычислительную стоимость подобных вычислительных экспериментов, исследования в этом направлении будут продолжены для проверки этой гипотезы.

6. ВЫВОДЫ

В данной работе рассмотрен алгоритм получения наиболее эффективной модели машинного обучения для прогнозирования геомагнитного индекса *Dst* путем постепенного отбрасывания входных признаков на основе следующих методов машинного обучения: линейная регрессия, градиентный бустинг, искусственная нейронная сеть типа многослойный персептрон. На основании полученных результатов можно сделать следующие основные выводы.

1) Отбор существенных входных признаков при прогнозировании значений геомагнитного индекса *Dst* методами машинного обучения позволяет кратно уменьшить количество используемых входных признаков без потери качества прогнозирования.

2) Отбираемые с помощью используемых адаптивных методов наиболее существенные входные признаки по своему физическому смыслу согласуются с существующими представлениями о влиянии различных физических величин на возмущение магнитосферы Земли. В качестве наименее существенной из рассмотренных была определена физическая величина *y*-компоненты межпланетного магнитного поля в системе GSM (которая и была включена в список входных признаков с целью проверки того, сможет ли алгоритм это обнаружить). В число наиболее существенных ожидаемо вошли значения самого прогнозируемого индекса *Dst* в широком диапазоне задержек, а также значения *Bz*-компоненты межпланетного магнитного поля в системе GSM и плотности протонов в солнечном ветре с малыми задержками относительно момента прогнозирования (от 0 до 3 ч).

3) Наилучшее качество прогнозирования на тестовом наборе данных показал алгоритм градиентного бустинга на полном наборе входных признаков или с отбором существенных входных признаков на базе этого же алгоритма. С увеличением горизонта прогноза качество прогноза для всех алгоритмов машинного обучения ощутимо падает.

4) Для логического завершения данного исследования необходимо поставить масштабный вычислительный эксперимент по отбору существенных входных признаков использованным в данной работе алгоритмом поочередного удаления признаков на базе многослойного персептрона. Полученные в таком эксперименте результаты по прогнозированию индекса *Dst* с помощью МСП на отобранном наборе признаков необходимо сравнить с результатами, полученными в настоящей работе с помощью градиентного бустинга и с помощью МСП.

5) Полученные в статье результаты по отбору существенных входных признаков справедливы для конкретной задачи прогнозирования геомагнитного индекса *Dst*. Однако сам по себе использованный алгоритм отбора является универсальным и может быть использован при решении других задач прогнозирования на основе многомерных временны́х рядов. В частности, представляет интерес применение данного алгоритма в процессе прогнозирования других геомагнитных индексов (*Kp*, *AE*, *AL*, *Ap* и др.) и сравнение множеств существенных признаков, отбираемых при прогнозировании разных геомагнитных индексов в одних и тех же временны́х диапазонах.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена по госбюджетной тематике Научно-исследовательского института ядерной физики им. Д.В. Скобельцына Московского государственного университета им. М.В. Ломоносова, темы 6.1 (01201255512) и 2.1 (115041410195).

СПИСОК ЛИТЕРАТУРЫ

– *Белов А.В., Виллорези Дж., Дорман Л.И. и др.* Влияние космической среды на функционирование искусственных спутников Земли // Геомагнетизм и аэрономия. Т. 44. № 4. С. 502–510. 2004.

– *Ермолаев Ю.И., Ермолаев М.Ю.* Солнечные и межпланетные источники геомагнитных бурь: Аспекты космической погоды // ГЕОФИЗИЧЕСКИЕ ПРОЦЕССЫ И БИОСФЕРА. Т. 8. № 1. С. 5–35. 2009.

– *Ефиторов А.О., Мягкова И.Н., Широкий В.Р., Доленко С.А.* Прогнозирование *Dst*-индекса, основанное на методах машинного обучения // Космич. исслед. Т. 56. № 6. С. 353–364. 2018.

– *Зорич В.А.* Многомерная геометрия, функции очень многих переменных и вероятность // Теория вероятностей и ее применения. Т. 59. Вып. 3. С. 436–451. 2014.

– *Лазутин Л.Л.* Мировые и полярные магнитные бури. М., МГУ. 214 с. 2012.

– *Мягкова И.Н., Шугай Ю.С., Веселовский И.С., Яковчук О.С.* Сравнительный анализ влияния рекуррентных высокоскоростных потоков солнечного ветра на радиационное состояние околоземного космического

пространства в апреле-июле 2010 года // Астрон. вестн. Т. 47. № 2. С. 141–155. 2013.

– *Мягкова И.Н., Широкий В.Р., Владимиров Р.Д., Баринов О.Г., Доленко С.А.* Прогнозирование значений геомагнитного индекса Dst при помощи адаптивных методов // Метеорология и гидрология. № 3. С. 38–46. 2021.

– *Широкий В.Р.* Сравнение нейросетевых моделей прогнозирования геомагнитного *Dst*-индекса на различных наборах данных и сравнение методов оценки качества работы моделей // XVII Всероссийская научно-техническая конференция "Нейроинформатика-2015" с международным участием. Сборник научных трудов. Ч. 2. М., НИЯУ МИФИ. С. 51–60. 2015.

– *Akasofu S.-I., S. Chapman S.* Solar-Terrestrial Physics. Clarendon Press, Oxford. 889 p. 1972.

– *Amata E., Pallocchia G., Consolini G. et al.* Comparison between three algorithms for *Dst* predictions over the 2003–2005 period //J Atmos Sol-Terr Phys. V. 70. P. 496–502. 2008.

– *Barkhatov N.A. et al.* Comparison of efficiency of artificial neural networks for forecasting the geomagnetic activity index *Dst* // Radiophysics and Quantum Electronics. V. 43. № 5. P. 347 –355. 2000.

– *Bortnik J., Chu X., Ma Q., Li W., Zhang X., Thorne R.M., Baker D.N.* Artificial Neural Networks for Determining Magnetospheric Conditions // Machine Learning Techniques for Space Weather. P. 279–300. 2018.

– *Breiman L., Friedman J.H., Olshen R., Stone C.* // Classification and Regression Trees. Wadsworth, Belmont, CA. 1983.

– *Breiman L.* Random Forests // Machine Learning. V. 45. P. 5–32. 2001

– *Burton R.K., McPherron R.L., Russell C.T.* An empirical relationship between interplanetary conditions and Dst // J. Geophys. Res. V. 80. P. 4204–4214. 1975.

– *Dolenko S.A., Orlov Yu.V., Persiantsev I.G., Shugai Ju.S.* Neural network algorithm for events forecasting and its application to space physics data // Lecture Notes in Computer Science. V. 3697. P. 527–532. 2005.

– *Friedman J.H.* Stochastic Gradient Boosting // Computational Statistics and Data Analysis. V. 38. № 4. P. 367–378. 2002.

– *Haykin S.* Neural Networks: A Comprehensive Foundation, 2nd ed. (Prentice Hall, 1998).

– *Kataoka R., Miyoshi Y.* Average profiles of the solar wind and outer radiation belt during the extreme flux enhancement of relativistic electrons at geosynchronous orbit // Ann. Geophys. V. 26. P. 1335–1339. 2008.

– *Lazzús J.A., Vega P., Rojas P., Salfate I.* Forecasting the *Dst* index using a swarm-optimizedneural network // Space Weather. V. 15. P. 1068–1089. 2017. https://doi.org/10.1002/2017SW001608

– *Lindsay G.M., Russell C.T., Luhmann J.G.* Predictability of *Dst* index based upon solar wind conditions monitored inside 1 AU // J. Geophys. Res. V. 104. № A5. P. 10335–10344. 1999.

– *Myagkova I., Shiroky V., Dolenko S.* Prediction of geomagnetic indexes with the help of artificial neural networks // E3S Web of Conferences, 20: art. 02011, 2017. https://doi.org/10.1051/e3sconf/20172002011

– *O'Brien T.P., McPherron R.L.* Forecasting the ring current index Dst in real time // J. Atmosph. and Sol.-Terrestr. Phys. V. 62. P. 1295–1299. 2000.

– *Pallochia G. et al.* Geomagnetic *Dst* -index forecast based on IMF data only // Ann. Geophys. V. 24. P. 989–999. 2006.

– *Patra S., Spencer E., Horton W., Sojka J.* Study of *Dst*/ring current recovery times using the WINDMI model //J. Geophys. Res. V.116. A02212. 2011. https://doi.org/10.1029/2010JA015824

– *Podladchikova T.V., Petrukovich A.A.* Extended geomagnetic storm forecast ahead of available solar wind measurements // Space Weather: The International J. Research and Applications. V. 10. CiteID S07001. 2012.

– *Pulkkinen T.* Space Weather: Terrestrial Perspective // Living Rev. Solar Phys. 4. 1. URL (cited on 18 September 2007): http://www.livingreviews.org/lrsp-2007-1. 2007.

– *Revallo M., Valach V., Hejda P., Bochníček J.* Modeling of CME and CIR driven geomagnetic storms by means of artificial neural networks // J. Atm. and Sol. Terr. Phys. V. 110. № 9. 2014.

– *Schrijver, Carolus J. et al.* Understanding space weather to shield society: A global road map 772 for 2015–2025 commissioned by COSPAR and ILWS // Adv. in Space Res. V. 55. P. 2745–2807. 2015.

– *Sugiura M.* Hourly values of equatorial *Dst* for the IGY // Ann. Int. Geophys. Pergamon Press, Oxford. V. 35. P. 9–45. 1964.

– *Wu J.-G., Lundstedt H.* Geomagnetic storm predictions from solar wind data with the use of dynamic neural networks // J. Geophys. Res. V. 102. № A7. P. 14255–14268. 1997.